

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

Taksonomia 22

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się

na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2014

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Eugeniusz Gatnar , Balance of payments statistics and external competitiveness of Poland.....	15
Andrzej Sokolowski, Magdalena Czaja , Efektywność metody k -średnich w zależności od separowalności grup.....	23
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw	30
Elżbieta Gołata , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
Aleksandra Łuczak, Feliks Wysocki , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów	49
Marek Walesiak , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej	60
Paweł Lula , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i>	69
Mariusz Kubus , Propozycja modyfikacji metody złagodzonego LASSO.....	77
Andrzej Bąk, Tomasz Bartłomowicz , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
Justyna Brzezińska , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki	104
Barbara Batóg, Jacek Batóg , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010	113
Małgorzata Markowska, Danuta Strahl , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
Kamila Migdał-Najman, Krzysztof Najman , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	131
Kamila Migdał-Najman, Krzysztof Najman , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	139
Beata Basiura, Anna Czapkiewicz , Badanie jakości klasyfikacji szeregów czasowych	148
Michał Trzęsiok , Wybrane metody identyfikacji obserwacji oddalonych.....	157

Grażyna Dehnel, Tomasz Klimanek , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
Michał Bernard Pietrzak, Justyna Wilk , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
Maciej Beręsewicz , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena	186
Marcin Szymkowiak, Tomasz Józefowski , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji	195
Marcin Pelka , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym	202
Małgorzata Machowska-Szewczyk , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
Justyna Wilk , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
Andrzej Dudek , Metody analizy skupień w klasyfikacji markerów map Google	229
Ewa Roszkowska , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW	237
Marcin Szymkowiak, Marek Witkowski , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
Bartłomiej Jefmański , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
Karolina Bartos , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych	266
Joanna Trzęsiok , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych	275
Beata Bal-Domańska , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
Beata Bieszk-Stolorz, Iwona Markowicz , Wpływ zasiłku na proces poszukiwania pracy	294
Marta Dziechciarz-Duda, Klaudia Przybysz , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
Tomasz Klimanek , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Wybrane metody analizy danych wzdluznych.....	321
Artur Zaborski , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych	330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

Katarzyna Wawrzyniak , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego	346
---	-----

Summaries

Eugeniusz Gatnar , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski	22
Andrzej Sokółowski, Magdalena Czaja , Cluster separability and the effectiveness of k -means method	29
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
Elżbieta Golata , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011	48
Aleksandra Łuczak, Feliks Wysocki , Determination of weights for features in problems of linear ordering of objects	59
Marek Walesiak , Reinforcing measurement scale for ordinal data in multivariate statistical analysis	68
Paweł Lula , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
Mariusz Kubus , The proposition of modification of the relaxed LASSO method.....	84
Andrzej Bąk, Tomasz Bartłomowicz , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
Justyna Brzezińska , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models	103
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
Barbara Batóg, Jacek Batóg , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity	120
Małgorzata Markowska, Danuta Strahl , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
Kamila Migdał Najman, Krzysztof Najman , Formal quality assessment of group structure mapping on the Kohonen's map	138
Kamila Migdał Najman, Krzysztof Najman , Graphical quality assessment of group structure mapping on the Kohonen's map	147
Beata Basiura, Anna Czapkiewicz , Validation of time series clustering	156
Michał Trzęsiok , Selected methods for outlier detection.....	166

Grażyna Dehnel, Tomasz Klimanek , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics	176
Michał Bernard Pietrzak, Justyna Wilk , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
Maciej Beręsewicz , An attempt to use different distance measures in the Generalized Petersen estimator	194
Marcin Szymkowiak, Tomasz Józefowski , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
Marcin Pelka , The ensemble conceptual clustering for symbolic data.....	209
Małgorzata Machowska-Szewczyk , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
Justyna Wilk , Problem of determining the number of clusters in taxonomic analysis of symbolic data	228
Andrzej Dudek , Clustering techniques for Google maps markers.....	236
Ewa Roszkowska , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure	247
Marcin Szymkowiak, Marek Witkowski , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
Bartłomiej Jefmański , The construction of fuzzy customer satisfaction indexes using R program.....	265
Karolina Bartos , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
Joanna Trzęsiok , Cluster analysis of countries with respect to fertility rate and other demographic factors	284
Beata Bal-Domańska , An attempt to identify major regional clusters and their convergence	293
Beata Bieszk-Stolorz, Iwona Markowicz , The influence of benefit on the job finding process	302
Marta Dziechciarz-Duda, Klaudia Przybysz , Education and labor market needs. Classification of university graduates	312
Tomasz Klimanek , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Selected methods for an analysis of longitudinal data.....	329
Artur Zaborski , The application of distance measures for ordinal data for aggregation individual preferences	337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market	345
Katarzyna Wawrzyniak , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows	355

Katarzyna Wawrzyniak

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

KLASYFIKACJA WOJEWÓDZTW WEDŁUG STOPNIA REALIZACJI PRIORYTETÓW STRATEGII ROZWOJU KRAJU 2007-2015 Z WYKORZYSTANIEM WARTOŚCI CENTRUM WIERSZOWEGO

Streszczenie: W artykule przedstawiono wyniki klasyfikacji województw według stopnia realizacji priorytetów strategicznych *Strategii rozwoju kraju 2007-2015* w 2011 r. Podstawą klasyfikacji był dwuelementowy proces diagnozowania, czyli porównanie prawidłowości zaobserwowanej z prawidłowością normatywną. Prawidłowość zaobserwowana została zdefiniowana jako wartości zmiennych ze złożonej macierzy znaczników, natomiast prawidłowość normatywna została określona na dwa sposoby – jako odpowiednio przekształcone wartości centrum wierszowego (norma średnia – wewnętrzna) oraz jako wartości docelowe (norma idealna – zewnętrzna).

Słowa kluczowe: diagnozowanie ilościowe, klasyfikacja, analiza korespondencji, złożona macierz znaczników, centrum wierszowe.

1. Wstęp

Artykuł jest kontynuacją badań prowadzonych przez autorkę dotyczących możliwości wykorzystania analizy korespondencji dla małych prób jako narzędzia diagnostycznego w mikro- i makroskali [Wawrzyniak 2011a, 2011b, 2012]. W dwóch pierwszych artykułach ostateczna diagnoza została sformułowana na podstawie końcowego wyniku analizy korespondencji, czyli rozrzutu punktów w przestrzeni dwuwymiarowej. Natomiast w trzecim artykule zasygnalizowano, że diagnozę uzyskaną na podstawie analizy rozrzutu punktów można uzupełnić poprzez wykorzystanie częstości brzegowych kolumn (średni profil wierszowy) jako norm wewnętrznych.

W obecnym artykule sformułowano hipotezę badawczą, że proces diagnozowania z wykorzystaniem klasycznej analizy korespondencji dla złożonej macierzy znaczników można zakończyć już na etapie wyznaczania macierzy częstości zaobserwowanych, w której częstości brzegowe kolumn i wierszy są odpowiednio śred-

nimi profilami wierszowymi (tzw. centrum wierszowe) i kolumnowymi (tzw. centrum kolumnowe). Do zweryfikowania powyższej hipotezy przyjęto następujące założenia:

- diagnozowanie jest to rozpoznanie wartościujące, w którym najpierw ma miejsce identyfikacja obecnego i prawidłowego stanu badanego zjawiska, a następnie – w celu sformułowania ostatecznej diagnozy – konieczna jest ocena przez porównanie obu tych stanów, aby stwierdzić, czy badane zjawisko przebiega zgodnie z normą – diagnoza pozytywna, czy też nie – diagnoza negatywna [Wawrzyniak 2007]¹,
- procedurę klasycznej analizy korespondencji dla złożonej macierzy znaczników można zdekomponować, uwzględniając diagnostyczny aspekt tej metody,
- diagnoza uzyskana na podstawie złożonej macierzy znaczników i wartości centrum wierszowego jest wystarczająca do dokonania klasyfikacji obiektów z punktu widzenia wybranych cech diagnostycznych.

Z ostatniego założenia wynika, że celem badania jest zaproponowanie sposobu klasyfikacji obiektów ze względu na poziom cech diagnostycznych z wykorzystaniem złożonej macierzy znaczników oraz wartości centrum wierszowego.

Weryfikacji hipotezy, a tym samym realizacji celu badania, dokonano na podstawie danych o wskaźnikach monitorujących stopień realizacji priorytetów strategicznych, przyjętych w *Strategii rozwoju kraju 2007-2015* dla województw w 2011 r., dostępnych na stronie Głównego Urzędu Statystycznego.

2. Charakterystyka prawidłowości zaobserwowanej i prawidłowości normatywnej

Wybór obszaru badawczego nie jest przypadkowy, gdyż do udowodnienia sformułowanej w artykule hipotezy konieczne jest zidentyfikowanie prawidłowości zaobserwowanej i normatywnej. W *Strategii rozwoju kraju 2007-2015 (SRK 2007-2015)*, która została przyjęta przez Radę Ministrów 29 listopada 2006 r., określono cele i priorytety strategiczne polityki rozwoju w perspektywie najbliższych lat. Zarówno stopień realizacji celów, jak i priorytetów jest monitorowany za pomocą określonych wskaźników, dla których w *SRK* przyjęto wartości docelowe dla 2010 i 2015 r. Rzeczywiste wartości wskaźników (prawidłowość zaobserwowana) monitorujących stopień realizacji priorytetów strategicznych wraz z wartościami docelowymi (prawidłowość normatywna) są publikowane na bieżąco przez Główny Urząd Statystyczny. W tabeli 1 podano docelowe wartości wskaźników monitorujących, które są dostępne dla województw. Ze względu na okres badania wartość

¹ Tak sformułowana definicja diagnozowania nawiązuje do definicji diagnozowania ekonometrycznego zaproponowanej przez J. Hozera [1989]. Według autora diagnozowanie ekonometryczne to proces składający się z następujących elementów: prawidłowości zaobserwowanej, prawidłowości normatywnej (normy), odchylenia od normy i tolerancji odchylenia od normy.

docelową wskaźników dla 2011 r. wyznaczono wykorzystując średni przyrost absolutny, obliczony na podstawie docelowych wartości tych wskaźników w latach 2010 i 2015.

Tabela 1. Wskaźniki monitorujące stopień realizacji priorytetów strategicznych *SRK 2007-2015* dostępne dla województw²

	Wartości docelowe (przyjęte w SRK)		Wartość docelowa (oszacowana)
	2010	2015	2011
Priorytet I. Wzrost konkurencyjności i innowacyjności gospodarki			
Nakłady ogółem na działalność badawczą i rozwojową w % PKB	1,5	2,0	1,6
Udział podmiotów gospodarczych w nakładach na działalność B+R w % – W_1	30	40	32
Priorytet II. Poprawa stanu infrastruktury technicznej i społecznej			
Ludność korzystająca z oczyszczalni ścieków w % ludności ogółem – W_2	75	85	77
Zgony z powodu chorób układu krążenia (liczba na 100 tys. ludności) – W_3	400	380	396
Priorytet III. Wzrost zatrudnienia i podniesienie jego jakości			
Wskaźnik zatrudnienia osób w wieku 15-64 lata (%) – W_4	57	62	58
Wskaźnik zatrudnienia kobiet w wieku 15-64 lata (%) – W_5	51	53	51,4
Wskaźnik zatrudnienia osób w wieku 55-64 lata (%) – W_6	31	37	32,2
Wskaźnik zatrudnienia osób niepełnosprawnych (%) – W_7	18	25	19,4
Priorytet IV. Budowa zintegrowanej wspólnoty społecznej i jej bezpieczeństwa			
Odsetek dzieci objętych wychowaniem przedszkolnym	50	55	51
Udział w wyborach parlamentarnych (% uprawnionych) – W_8	43	50	44,4
Wskaźnik wykrywalności sprawców przestępstw (%)	60	65	61
Priorytet V. Rozwój obszarów wiejskich			
Wskaźnik zatrudnienia osób w wieku 15 lat i więcej na wsi (%) – W_9	51	54	51,6
Przeciętna powierzchnia gospodarstwa indywidualnego wynosząca powyżej 1 ha użytków rolnych (ha użytków rolnych) – W_{10}	8,5	10,0	8,8
Odsetek dzieci w wieku 3-5 lat objętych wychowaniem przedszkolnym w placówkach na wsi	25	35	27
Priorytet VI. Rozwój regionalny i podniesienie spójności terytorialnej			
Wskaźnik urbanizacji – W_{11}	62	65	62,6

Źródło: opracowanie własne na podstawie http://www.stat.gov.pl/gus/wskazniki_monitorujace_PLK_HTML.htm (17.04.2013).

² W dalszej części artykułu wykorzystano tylko te wskaźniki, przy których w tabeli 1 znajduje się symbol W_i ($i = 1, 2, \dots, 11$). Trzy wskaźniki pominięto, gdyż dla wszystkich województw ich wartości w 2011 r. albo przekroczyły, albo były poniżej wartości docelowej. To z kolei spowodowało, że częstości brzegowe kolumn utworzone na podstawie złożonej macierzy znaczników były równe zero i nie można było zastosować klasycznej analizy korespondencji.

3. Klasyczna analiza korespondencji dla złożonej macierzy znaczników a proces diagnozowania

W artykule badaną zbiorowość tworzy 16 województw scharakteryzowanych za pomocą 11 wskaźników monitorujących, z których tylko jeden (W_3) jest destymulantą, a pozostałe to stymulanty. W przypadku małej zbiorowości ($n = 16$) zastosowanie klasycznej analizy korespondencji wymaga odpowiedniego przygotowania danych wyjściowych albo w postaci złożonej macierzy znaczników, albo poprzez wykorzystanie metody podwajania obserwacji [Stanimir 2005, s. 108-110]. Ze

Tabela 2. Dekompozycja klasycznej analizy korespondencji dla złożonej macierzy znaczników z uwzględnieniem aspektu diagnostycznego

Lp.	Etapy postępowania w analizie korespondencji	Aspekt diagnostyczny
1	Budowa złożonej macierzy znaczników	Dwuelementowy proces diagnozowania polegający na przekształceniu każdej zmiennej diagnostycznej na dwie nowo utworzone zmienne zero-jedynkowe poprzez porównanie rzeczywistej wartości tej zmiennej z normą $W_i \rightarrow X_{ip}, X_{in}$ X_{ip} – dla stymulant $x_{ip} = \begin{cases} 1 & \text{dla } w_i \geq \text{norma} \\ 0 & \text{dla } w_i < \text{norma} \end{cases}$ – dla destymulant $x_{ip} = \begin{cases} 1 & \text{dla } w_i \leq \text{norma} \\ 0 & \text{dla } w_i > \text{norma} \end{cases}$ X_{in} $x_{in} = \begin{cases} 1 & \text{gdy } x_{ip} = 0 \\ 0 & \text{gdy } x_{ip} = 1 \end{cases}$
2	Wyznaczenie macierzy częstości zaobserwowanych, w której częstości brzegowe kolumn (e) i wierszy (r) są odpowiednio średnimi profilami wierszowymi (centrum wierszowe) i kolumnowymi (centrum kolumnowe)	Wartości centrum wierszowego (e) i centrum kolumnowego (r) można uznać za prawidłowości normatywne odpowiednio dla poszczególnych profili wierszowych zapisanych w macierzy (R) oraz profili kolumnowych zapisanych w macierzy (C) – wartości zapisane w macierzy R i C można potraktować jako prawidłowości zaobserwowane
3	Wyznaczenie rzeczywistej przestrzeni rzutowania	
4	Obliczenie współrzędnych kategorii zmiennych z wykorzystaniem rozkładu macierzy według wartości osobliwych	
5	Ocena jakości odwzorowania powiązań pomiędzy kategoriami zmiennych z rzeczywistej przestrzeni rzutowania w przestrzeń o niższym wymiarze (dwuwymiarowym)	
6	Graficzna prezentacja wyników w przestrzeni dwuwymiarowej	
7	Interpretacja rozrzutu punktów w przestrzeni dwuwymiarowej	W tym etapie możliwe jest zdiagnozowanie słabych i mocnych stron badanych obiektów pod względem poziomu tych zmiennych, które w istotny sposób odróżnią te obiekty od pozostałych

Źródło: opracowanie własne z wykorzystaniem opisu algorytmu klasycznej analizy korespondencji z prac [Gatnar, Walesiak (red.) 2004, s. 285-294; Stanimir 2005, s. 21-29].

Tabela 3. Złożona macierz znaczników wraz z wartościami centrum wierszowego

Woje- wództwo*	W ₁		W ₂		W ₃		W ₄		W ₅		W ₆		W ₇		W ₈		W ₉		W ₁₀		W ₁₁		Suma (n _i)
	X _{1p}	X _{1n}	X _{2p}	X _{2n}	X _{3p}	X _{3n}	X _{4p}	X _{4n}	X _{5p}	X _{5n}	X _{6p}	X _{6n}	X _{7p}	X _{7n}	X _{8p}	X _{8n}	X _{9p}	X _{9n}	X _{10p}	X _{10n}	X _{11p}	X _{11n}	
DL	1	0	0	1	0	1	0	1	1	0	1	0	0	1	1	0	0	1	1	0	1	0	11
KP	0	1	0	1	0	1	0	1	0	1	1	0	1	0	1	0	0	1	1	0	0	1	11
LL	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0	0	1	0	1	11
LB	0	1	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	1	1	0	1	11
ŁD	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1	0	11
MP	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	11
MZ	0	1	0	1	0	1	1	0	1	0	1	0	0	1	1	0	1	0	0	1	1	0	11
OP	1	0	0	1	0	1	1	0	0	1	1	0	1	0	0	1	0	1	1	1	0	0	11
PK	1	0	0	1	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0	1	11
PL	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	0	1	1	0	0	1	11
PM	1	0	1	0	1	0	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1	0	11
ŚL	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	11
ŚW	0	1	0	1	0	1	1	0	1	0	1	0	1	0	0	1	1	0	0	1	0	1	11
WM	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1	11
WP	0	1	0	1	1	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0	0	1	11
ZP	0	1	1	0	0	1	0	1	0	1	1	0	0	1	1	0	0	1	1	0	1	0	11
Suma (n _j)	4	12	2	14	3	13	9	7	8	8	14	2	11	5	12	4	5	11	9	7	7	9	176(n)
Wartości centrum wierszowego ($\frac{\sum_j x_{ij}}{n}$)																							
0,023 0,068 0,011 0,080 0,017 0,074 0,051 0,040 0,046 0,046 0,080 0,011 0,063 0,028 0,068 0,023 0,028 0,063 0,051 0,040 0,040 0,051 1,000																							

* Zamiast pełnych nazw województw podano ich dwuliterowe skróty. Skróty te będą wykorzystywane w kolejnych tabelach w dalszej części artykułu.

Źródło: obliczenia własne na podstawie danych ze strony GUS http://www.stat.gov.pl/gus/wskazniki_monitorujace_PLK_HTML.htm (17.04.2013).

względu na cel badania zdecydowano się na wykorzystanie złożonej macierzy znaczników, gdyż przy jej budowie można zaobserwować dwuelementowy proces diagnozowania. W tabeli 2 zaprezentowano dekompozycję klasycznej analizy korespondencji dla złożonej macierzy znaczników, wskazując te etapy, w których ma miejsce proces diagnozowania albo występują elementy tego procesu.

W tabeli 3 przedstawiono złożoną macierz znaczników o wymiarach 16×22 , gdzie 16 oznacza liczbę województw, a 22 oznacza liczbę nowych zmiennych zero-jedynkowych utworzonych zgodnie z zasadą przedstawioną w tabeli 2 w etapie pierwszym. W ostatnim wierszu tabeli 3 podane zostały wartości centrum wierszowego. Elementy przedstawione w tabeli 3 wykorzystano w procesie klasyfikacji województw według stopnia realizacji priorytetów strategicznych *SRK 2007-2015* w 2011 r.

4. Klasyfikacja województw z wykorzystaniem wartości centrum wierszowego

Do przeprowadzenia klasyfikacji województw według stopnia realizacji priorytetów strategicznych *SRK 2007-2015* w 2011 r. z wykorzystaniem wartości centrum wierszowego z tabeli 3 wybrano tylko te kolumny, które odpowiadały nowo utworzonym zmiennym o symbolu X_{ip} . Dla tych zmiennych wartość 1 oznacza, że w i -tym województwie wartość k -tego wskaźnika monitorującego jest zgodna z normą, natomiast wartość 0 oznacza brak takiej zgodności. Wartości te potraktowano jako prawidłowości zaobserwowane dla poszczególnych województw. Jako prawidłowości normatywne przyjęto odpowiadające tym kolumnom wartości centrum wierszowego. Aby zapewnić porównywalność obu prawidłowości, wartości centrum wierszowego przekształcono na zmienną zero-jedynkową według zasady: jeżeli suma jedynek w danej kolumnie była przynajmniej równa 8, to wartości centrum wierszowego przyporządkowano liczbę 1, w przeciwnym przypadku liczbę 0. Przyjęcie tego założenia oznacza, że w całej zbiorowości przynajmniej 50% województw uzyskało poziom wskaźnika monitorującego zgodny z normą i cała zbiorowość z punktu widzenia tego wskaźnika została oceniona pozytywnie. Przekształcone wartości centrum wierszowego uznano za normę średnią wynikającą z rozkładu zmiennych X_{ip} w całej zbiorowości. W tabeli 4 zaprezentowano wartości zmiennych X_{ip} dla poszczególnych województw oraz normy średnie i docelowe³.

Informacje zamieszczone w tabeli 4 umożliwiły przeprowadzenie dwuelementowego procesu diagnozowania (porównanie prawidłowości zaobserwowanej z prawidłowością normatywną, czyli obliczenie różnic między wartością zmiennej

³ Wartości norm docelowych przyjęto dla każdej zmiennej X_{ip} na poziomie 1, czyli przyjęto sytuację idealną, w której poziom wszystkich wskaźników monitorujących we wszystkich województwach jest zgodny z normą.

Tabela 4. Wartości zmiennych X_{ip} oraz norm średnich i docelowych

Województwa	X_{1p}	X_{2p}	X_{3p}	X_{4p}	X_{5p}	X_{6p}	X_{7p}	X_{8p}	X_{9p}	X_{10p}	X_{11p}
DL	1	0	0	0	1	1	0	1	0	1	1
KP	0	0	0	0	0	1	1	1	0	1	0
LL	0	0	0	1	1	1	1	1	1	0	0
LB	0	0	0	0	1	1	1	0	0	1	1
ŁD	0	0	0	1	1	1	1	1	1	0	1
MP	0	0	0	1	1	1	1	1	0	0	0
MZ	0	0	0	1	1	1	0	1	1	0	1
OP	1	0	0	1	0	1	1	0	0	1	0
PK	1	0	0	0	0	1	1	1	0	0	0
PL	0	0	0	1	1	1	1	1	0	1	0
PM	1	1	1	1	0	1	1	1	0	1	1
ŚL	0	0	0	0	0	0	0	1	0	0	1
ŚW	0	0	0	1	1	1	1	0	1	0	0
WM	0	0	1	0	0	0	0	0	0	1	0
WP	0	0	1	1	0	1	1	0	1	1	0
ZP	0	1	0	0	0	1	0	1	0	1	1
Suma 1 w kolumnie	4	2	3	9	8	14	11	12	5	9	7
Wartości centrum wierszowego	0,023	0,011	0,017	0,051	0,045	0,080	0,063	0,068	0,028	0,051	0,040
Norma średnia	0	0	0	1	1	1	1	1	0	1	0
Norma docelowa	1	1	1	1	1	1	1	1	1	1	1

Źródło: opracowanie własne na podstawie tabeli 3.

X_{ip} a wartością normy (średniej lub docelowej), którego wyniki w skróconej formie zostały zaprezentowane w tabeli 5. Wartości z tej tabeli dla normy średniej należy interpretować następująco:

- wartość -1 (różnica między $x_{ip} = 0$ i normą = 1) oznacza, że poziom danego wskaźnika dla i -tego województwa jest poniżej normy, czyli województwo nie należy do grupy województw ocenionych pozytywnie ze względu na poziom tego wskaźnika,
- wartość 0 (różnica między $x_{ip} = 0$ i normą = 0 lub między $x_{ip} = 1$ i normą = 1) oznacza, że poziom danego wskaźnika dla i -tego województwa jest zgodny z normą, przy czym w przypadku pierwszym województwo należy do grupy województw ocenionych negatywnie ze względu na poziom tego wskaźnika⁴, a w przypadku drugim – do grupy województw ocenionych pozytywnie,

⁴ Do wyznaczenia liczby wskaźników spełniających przypadek pierwszy – niekorzystny z punktu widzenia diagnozy – można wykorzystać różnicę obliczoną między liczbą wskaźników o wartościach powyżej i zgodnych z normą średnią a liczbą wskaźników o wartościach zgodnych z normą docelową (np. w województwie pomorskim (PM) z taką sytuacją mieliśmy do czynienia tylko dla jednego wskaźnika, a w województwie kujawsko-pomorskim (KP) aż z pięcioma – porównaj tabela 6).

- wartość 1 (różnica między $x_{ip} = 1$ i normą = 0) oznacza, że poziom danego wskaźnika dla i -tego województwa jest powyżej normy, czyli województwo nie należy do grupy województw ocenionych negatywnie ze względu na poziom tego wskaźnika, ale należy do grupy województw ocenionych pozytywnie.

Tabela 5. Wyniki dwuelementowego procesu diagnozowania

Województwa	X _{1p}	X _{2p}	X _{3p}	X _{4p}	X _{5p}	X _{6p}	X _{7p}	X _{8p}	X _{9p}	X _{10p}	X _{11p}
Porównanie prawidłowości zaobserwowanej z normą średnią (wewnętrzną)											
DL	1	0	0	-1	0	0	-1	0	0	0	1
KP	0	0	0	-1	-1	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
WP	0	0	1	0	-1	0	0	0	1	0	0
ZP	0	1	0	-1	-1	0	-1	0	0	0	1
Porównanie prawidłowości zaobserwowanej z normą docelową (zewnętrzną)											
DL	0	-1	-1	-1	0	0	-1	0	-1	0	0
KP	-1	-1	-1	-1	-1	0	0	0	-1	0	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
WP	-1	-1	0	0	-1	0	0	0	0	0	-1
ZP	-1	0	-1	-1	-1	0	-1	0	-1	0	0

Źródło: obliczenia własne na podstawie tabeli 4.

Tabela 6. Klasyfikacja województw według stopnia realizacji priorytetów strategicznych SRK 2007-2015 w 2011 r.

Lp.	Województwo	Norma średnia		
		liczba wskaźników		
		powyżej normy	zgodnych z normą	poniżej normy
1	PM	4	6	1
2	ŁD	2	8	1
3	WP	2	8	1
4	DL	2	7	2
5	MZ	2	7	2
6	ZP	2	6	3
7	LL	1	9	1
8	LB	1	8	2
9	OP	1	8	2
10	ŚW	1	8	2
11	PK	1	7	3
12	ŚL	1	5	5
13	WM	1	5	5
14	PL	0	11	0
15	MP	0	10	1
16	KP	0	9	2

Lp.	Województwo	Norma docelowa	
		liczba wskaźników	
		zgodnych z normą	poniżej normy
1	PM	9	2
2	ŁD	7	4
3	WP	7	4
4	DL	6	5
5	LL	6	5
6	MZ	6	5
7	PL	6	5
8	LB	5	6
9	MP	5	6
10	OP	5	6
11	ŚW	5	6
12	ZP	5	6
13	KP	4	7
14	PK	4	7
15	ŚL	2	9
16	WM	2	9

Źródło: opracowanie własne na podstawie tabeli 5.

W przypadku normy docelowej (wzorzec idealny) w tabeli 5 występują tylko dwie wartości -1 (różnica obliczana tak jak dla normy średniej) i 0 (różnica między $x_{ip} = 1$ i normą = 1). Brak w tej części tabeli wartości 1 wynika z tego, że przy tworzeniu normy docelowej założono, że we wszystkich województwach poziom danego wskaźnika jest zgodny z normą i nie może już być województwa lepszego.

W tabeli 6 przedstawiono wyniki klasyfikacji województw według stopnia realizacji priorytetów strategicznych *SRK 2007-2015* w 2011 r. w porównaniu z normą⁵.

5. Podsumowanie

W artykule wykazano, że jeżeli celem badania jest dokonanie klasyfikacji obiektów ze względu na poziom zmiennych diagnostycznych, to procedurę obliczeniową klasycznej analizy korespondencji można zakończyć już po utworzeniu złożonej macierzy znaczników oraz wyznaczeniu częstości brzegowych kolumn, czyli wartości centrum wierszowego. Informacje zawarte w odpowiednio zbudowanej złożonej macierzy znaczników należy potraktować wówczas jako prawidłowości zaobserwowane, a odpowiednio przekształcone wartości centrum wierszowego – jako prawidłowości normatywne (norma średnia). Porównanie tych dwóch elementów jest wystarczające do przeprowadzenia klasyfikacji obiektów. Istotną zaletą przedstawionej propozycji jest możliwość wprowadzenia normy zewnętrznej (docelowej, idealnej), która daje podstawy do kolejnej klasyfikacji oraz pozwala wskazać liczbę zmiennych, których poziom jest zgodny z normą docelową (idealną). Przeprowadzenie pełnej procedury klasycznej analizy korespondencji jest zasadne wówczas, gdy celem badania jest pogrupowanie obiektów, aby zdiagnozować słabe i mocne strony obiektów pod względem poziomu tych zmiennych, które w istotny sposób odróżnią te obiekty od pozostałych.

Literatura

- Hozer J. (1989), *Funkcja diagnostyczna modeli ekonometrycznych*, „Wiadomości Statystyczne”, nr 2, s. 13-15.
- Gainar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.
- Stanimir A. (2005), *Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.
- Strategia rozwoju kraju 2007-2015* (2006), Ministerstwo Rozwoju Regionalnego, Warszawa, listopad, <http://www.mrr.gov.pl>.

⁵ Dla normy średniej im większa liczba wskaźników powyżej normy i zgodnych z normą, a mniejsza liczba wskaźników poniżej normy, tym wyższe miejsce w rankingu. W przypadku normy docelowej miejsce w rankingu było tym wyższe, im więcej było wskaźników zgodnych z normą, a im mniej poniżej normy.

- Wawrzyniak K. (2007), *Diagnozowanie ilościowe procesów i obiektów gospodarczych – podstawowe pojęcia*, [w:] J. Hozer (red.), *Metody ilościowe w ekonomii*, Zeszyty Naukowe US nr 450, Wydawnictwo Naukowe US, Szczecin, s. 647-659.
- Wawrzyniak K. (2011a), *Analiza korespondencji jako narzędzie diagnostyczne w makroskali*, [w:] J. Dziechciarz (red.), *Zastosowanie metod ilościowych*, Prace Naukowe UE we Wrocławiu nr 163, *Ekonometria* 30, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 19-27.
- Wawrzyniak K. (2011b), *Diagnoza sytuacji finansowo-ekonomicznej spółek giełdowych z wykorzystaniem klasycznej analizy korespondencji*, „*Oeconomica*”, *Folia Pomer. Univ. Technol. Stetin.* nr 285 (62), Wydawnictwo ZUT w Szczecinie, Szczecin, s. 105-116.
- Wawrzyniak K. (2012), *Ocena stopnia realizacji celu głównego Strategii rozwoju kraju według województw*, „*Wiadomości Statystyczne*” nr 12, s. 36-52.

CLASSIFICATION OF VOIVODESHIPS ACCORDING TO THE LEVEL OF THE REALIZATION OF PRIORITIES OF THE NATIONAL DEVELOPMENT STRATEGY 2007-2015 WITH USING THE VALUES OF CENTROID OF THE ROWS

Summary: The paper presents the results of the classification of voivodeships according to the level of the realization of priorities of *the National Development Strategy 2007-2015* in 2011. The classification was based on the two-element process of diagnosing which involves comparing the observed regularity with normative regularity. The observed regularity was defined as values of variables from the multiple indicator matrix. The normative regularity was identified in two ways – as the properly transformed values of centroid of the rows (average norm – internal) and as the target values (ideal norm – external).

Keywords: quantitative diagnosis, classification, correspondence analysis, multiple indicator matrix, centroid of the rows.